Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Multiresolution alignment for multiple unsynchronized audio sequences using sequential Monte Carlo samplers

Dogac Basaran [a,*], Ali Taylan Cemgil [b], Emin Anarim [c]

[a] Signal and Image Processing Department, Telecom-Paristech University, 46 Rue Barrault, Paris, France
[b] Computer Engineering Department, Bogazici University, 34342 Bebek, Istanbul, Turkey
[c] Electrical and Electronics Engineering Department, Bogazici University, 34342 Bebek, Istanbul, Turkey

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Multiple audio alignment
Multiresolution alignment
Audio fingerprint
Bayesian inference
Sequential Monte Carlo samplers
Sequential alignment

ABSTRACT

It is increasingly more common that an occasion is recorded by multiple individuals with the proliferation of recording devices such as smart phones. When properly aligned, these recordings may provide several audio and visual perspectives to a scene which leads to several applications in restoring, remastering and remixing frameworks in various fields. In this work, we propose a multiresolution alignment algorithm for aligning multiple unsynchronized audio sequences using Sequential Monte Carlo samplers. We employ a model based approach and a score function analogous to similarity based methods. The optimum alignments are obtained in a course to fine structure with multiresolution sampling and a heuristic sequential search method. The proposed method is evaluated with a real-life dataset from Jiku Mobile Video Datasets. The simulation results suggest that our method is competitive with the baseline methods in terms of accuracy with suitable choice of parameters.

1. Introduction

With the proliferation of recording devices and applications for user generated content sharing, an increasing number of people regularly capture audio and video in special occasions like concerts, conferences and sports competitions. As a consequence, a single event can be simultaneously recorded by multiple individuals (such as using smart phones) creating wide coverage, multiple visual and listening perspectives to a scene. If privacy is not a concern, these user generated multimedia (audio/video) data are typically made accessible through social media sharing sites however in unorganized form. Temporally aligning and combining such data could lead to a wide range of applications. In [1], audience generated video clips from a concert event are aligned using audio features to obtain a full-clip of one song. In [2], over 700 YouTube videos related to a U.S. presidential inauguration are used to restore the U.S. President's speech. An application of automatic video remixing can be found in [3], where the system automatically creates remixes from videos recorded by mobile devices. Multimedia alignment also has applications in forensics field [4] and lately, 360-degrees video creation became very popular [5]. There are also commercially available video synchronization tools such as PluralEyes [6] and DualEyes.

1.1. Problem statement

We describe the problem setup with the following example. Imagine you attend a concert of your favorite band. During the concert, several people from the audience record some parts of the concert with their smart phones, from different perspectives and independent from each other. These recordings do not necessarily contain the same audio/visual content i.e., recordings of entirely different songs, and probably none of recordings cover the entire concert. The quality of each recording device might differ depending on the hardware, compression distortions as well as the environmental contaminating noise. In this setting, the aim is to temporally synchronize these multimedia recordings relative to each other on a common timeline utilizing the audio content.

We formally define the alignment problem as following. There is a dataset of K user generated, unsynchronized recordings denoted as $\mathbf{x} = \{\mathbf{x}_k\}_{k=1}^{K}$. We denote the offsets of sequences referenced to universal (generic) time line as $\mathbf{r} = \{r_k\}_{k=1}^{K}$. If a pair of sequences $(\mathbf{x}_i, \mathbf{x}_j)$ are overlapping on the universal time line, we call these sequences as *connected* (*connected* and *overlapping* are interchangeably used in the text). The set of connected sequences form a cluster $\mathbf{C} = \{\mathcal{C}_m\}_{m=1}^{M}$ i.e., x_i and x_j are connected, x_j and x_l are connected then they form the cluster $C_m = \{x_i, x_j, x_l\}$. There are $1 \le M \le K$ disjoint (not connected) clusters. The alignment prob-

[*] Corresponding author.
E-mail addresses: dogac.basaran@telecom-paristech.fr (D. Basaran), taylan.cemgil@boun.edu.tr (A.T. Cemgil), anarim@boun.edu.tr (E. Anarim).

https://doi.org/10.1016/j.dsp.2017.10.024
1051-2004/© 2017 Elsevier Inc. All rights reserved.

Please cite this article in press as: D. Basaran et al., Multiresolution alignment for multiple unsynchronized audio sequences using sequential Monte Carlo samplers, Digit. Signal Process. (2017), https://doi.org/10.1016/j.dsp.2017.10.024

lem is then to determine each connected pair of sequences $(\mathbf{x}_i, \mathbf{x}_j)$ in the dataset, each disjoint cluster and further determine the relative time offsets $\Delta\mathbf{r} = \{\Delta r_{ij} = |r_i - r_j|\}_{i,j}$ for all connected pairs $(\mathbf{x}_i, \mathbf{x}_j)$.

Note that the sequences $\mathbf{x}$ are usually discrete time–frequency representations of raw audio signals such as short-time Fourier transform (STFT). Then each sequence is represented as $\mathbf{x}_k = \{x_{fn}\}_{f=1,n=1}^{F,N_k}$ where f denotes the frequency band index, n denotes the frame index, F denotes the number of frequency bands and N_k denotes the length of the sequence $\mathbf{x}_k$ (in frames). In this manner, the offsets of sequences $\mathbf{r}$ are also treated in frames instead of in seconds.

1.2. Related work

In state-of-the-art, audio alignment problem is tackled in a twofold manner: First the audio signals are represented with robust features against various types of noise and distortions, then search algorithms are employed over those representations to find the best offset setting of sequences usually utilizing exact hash (fingerprint) matches, similarity or cost functions.

Audio representations in existing methods mostly involve audio fingerprints [7,8,1,9–14], transform domain (time–frequency) features such as chroma [15,16], spectral flatness [17], spectral energy [18] and audio onsets [19,20].

Although audio fingerprints, as compact signature representations of audio, were originally developed for music identification services, query-by-example based indexing schemes for audio identification are utilized for audio alignment purposes in [1,9,11,12,14,20,21]. These methods are usually require low computational time as an advantage. The usual approach requires to extract hashes (fingerprints) from audio sequences and the number of exact hash matches between sequences is used to determine if there is a match. Then for each matching pair, the relative offset is computed[1] and connected sequences are grouped to form clusters [14]. A similar query-by-example method is proposed in [15] where audio chroma features are used instead of binary fingerprints. A descriptor is defined as a classifier to find matching sequences.

One major drawback with the fingerprinting approach is that in real-life conditions, two matching audio sequences may have very few or no exact matching fingerprints under some distortions and low SNR conditions. Besides that, a matching decision between two audio sequences is achieved via thresholding the number of exact fingerprint matches which is hard to set for a global solution.

A straightforward way to tackle the alignment problem is utilizing similarity measures such as cross-correlation [9,17,13] or Hamming distance [13]. Methods utilizing such similarity measures usually apply matching for each pair of sequences by using thresholding methods. Then for the matching sequences, the relative offset with highest value (most similar) is accepted as the best estimate.

In [13], the data set is pre-classified into classes such as silence, music, speech and noise, before aligning with cross-correlation.

These measures are easy to implement, fast and robust however they have two major drawbacks. First of all, similar to fingerprinting based approaches, matching sequences are estimated using ad-hoc thresholds that depends highly on data. Secondly, these methods do not provide finding the amount of similarity of a sequence against a cluster of pre-aligned sequences i.e., each pair of sequences have to visited to find matchings. To overcome this problem, a greedy merging method is applied in [9] to form clusters of aligned sequences so that another sequence can be matched with the cluster. In [18], scoring functions similar to cross-correlation and Hamming distance are proposed that solves how to align a sequence against a cluster and how to determine matching sequences automatically.

In this paper, we propose a course to fine structure, multiresolution audio alignment scheme that can be applied to an arbitrary number of sequences ($K > 2$) using Sequential Monte Carlo (SMC) samplers. Note that the initial phase of this work is presented in [22] where the multiresolution scheme is considered for aligning pairs of sequences. The main intuition of the multiresolution alignment is that aligning the sequences in a courser level with a low computational effort and sequentially refining the estimated alignment.

Here, we extend the idea of multiresolution alignment via SMC samplers to a multiple audio alignment setting where we draw samples (alignment estimates) from a multidimensional, multimodal likelihood surface defined in [18] that penalizes the alignment of K sequences. SMC sampler particularly fit to the multiresolution setting because it samples the target surface sequentially through a sequence of intermediate distributions each distribution being known up to a normalizing constant [23]. The SMC method is based on sequential importance sampling [24–26] and it is flexible in design i.e., intermediate distributions can have different resolutions.

The main contributions of this work can be listed as follows:

- To our knowledge, this is the first study that proposes a multiresolution alignment method for aligning *multiple* user generated multimedia content.
- A SMC sampler mechanism is defined for multiple audio alignment setting that is able to sample from the likelihood of any alignment setting of K sequences.

The proposed method is evaluated with a real-life dataset from Jiku Mobile Video Dataset [27]. The results are compared to a fingerprinting based baseline method in terms of well-known metrics. The accompanying software architecture and impact are given in the joint manuscript [28] and the software is available online.[2]

The rest of the paper is organized as follows: In Section 2, we briefly explain the score function in [18] and the probabilistic model that it is derived from as well as the interpretation in a multiresolution setup. Then in Section 3, the multiresolution alignment using SMC samplers is explained extending to a multiple audio alignment setting. In Section 4, the experimental setup, implementation issues, the evaluation results and discussion are given. Then in Section 5, conclusions are given and some future directions are discussed.

2. Model based approach

In this section, we explain the probabilistic modeling approach [18,29] and how it can be used in a multiresolution fashion in multiple audio alignment setting.

2.1. Model

In addition to the definitions given in Section 1.1, we further define a random variable $\boldsymbol{\lambda} = \{\lambda_{f\tau}\}_{f=1,\tau=1}^{F,T}$ where f is the frequency bin index, τ is the frame index (on the generic time line) and T is the length of the sequence in frames. Here, $\boldsymbol{\lambda}$ denotes an unobserved parameter sequence but common with the observed sequences. The central theme of the model is as follows: *Given*

[1] Most of the audio fingerprints have the time-stamp information hence the relative time difference is able to be computed.

[2] https://github.com/dogacbasaran/Multiple-Audio-Alignment.

Please cite this article in press as: D. Basaran et al., Multiresolution alignment for multiple unsynchronized audio sequences using sequential Monte Carlo samplers, Digit. Signal Process. (2017), https://doi.org/10.1016/j.dsp.2017.10.024